Normal/ Gaussain Distribution in Statistics

It is a bell-shaped curve with continuous probability distribution where most data points cluster around the mean, and the probabilities decrease symmetrically as you move away from the center.

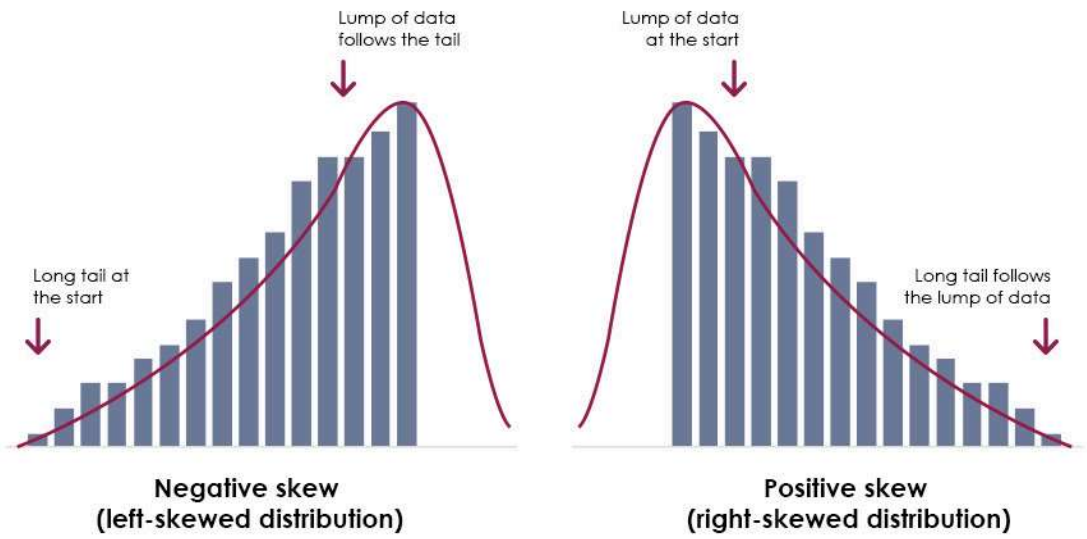
X ~ N (μ,σ)

Why Normal Distribution is Important?

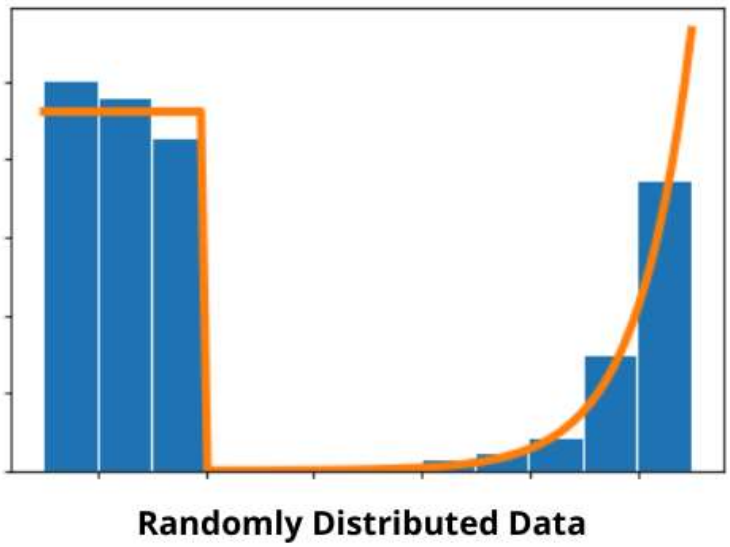
Data can be "distributed" (spread out) in different ways.

It can be spread out:

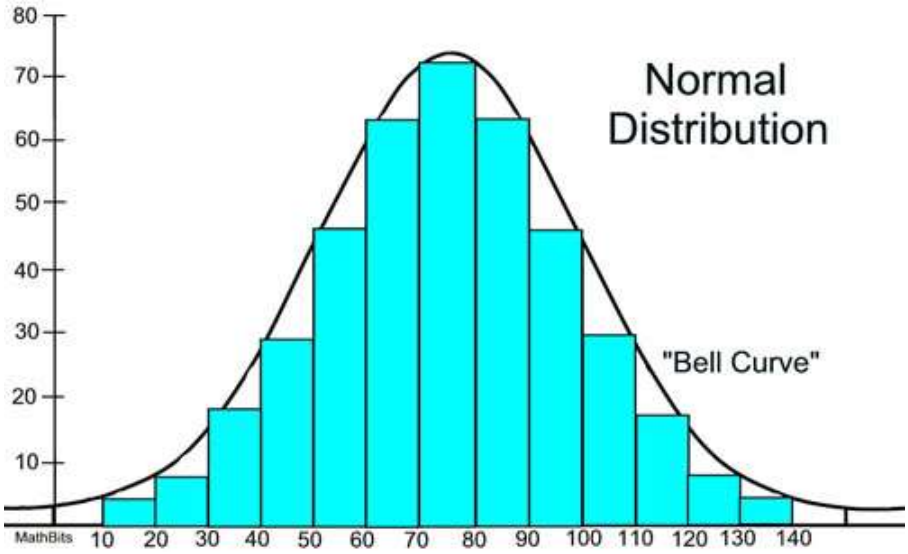
more on the left Or more on the right



Or it can be all random data



But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a **"Normal Distribution"** like this:



The black curve is called a Normal Distribution.

The blue histogram shows some data that closely follows it, but not perfectly (which is usually the case with data).

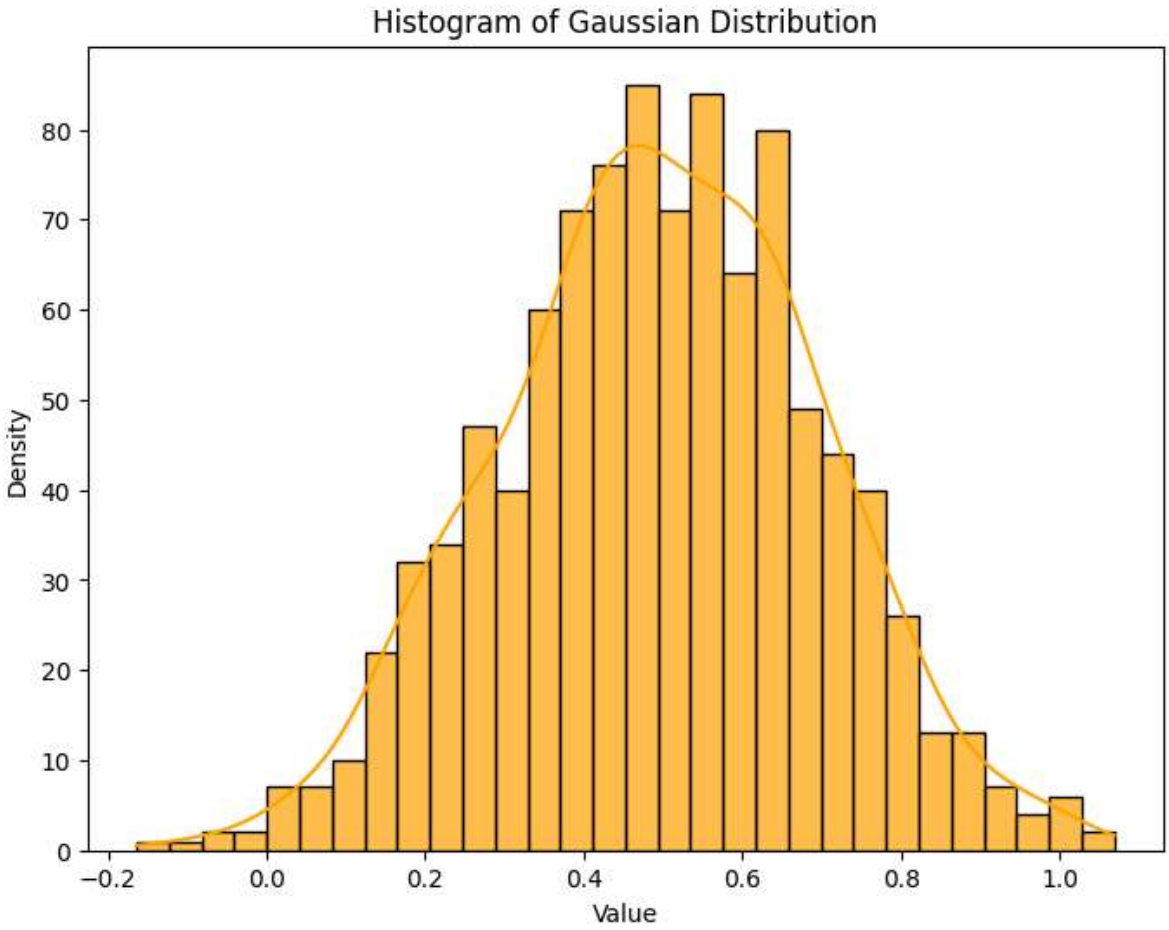
Many things closely follow a Normal Distribution:

- weights/ heights/ blood pressure of people
- size of things produced by machines
- pizza delivery time
- errors in measurements or experimental data.
- stock market returns over a long period of time.
- SAT/JEE/GRE scores

In [11]:
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

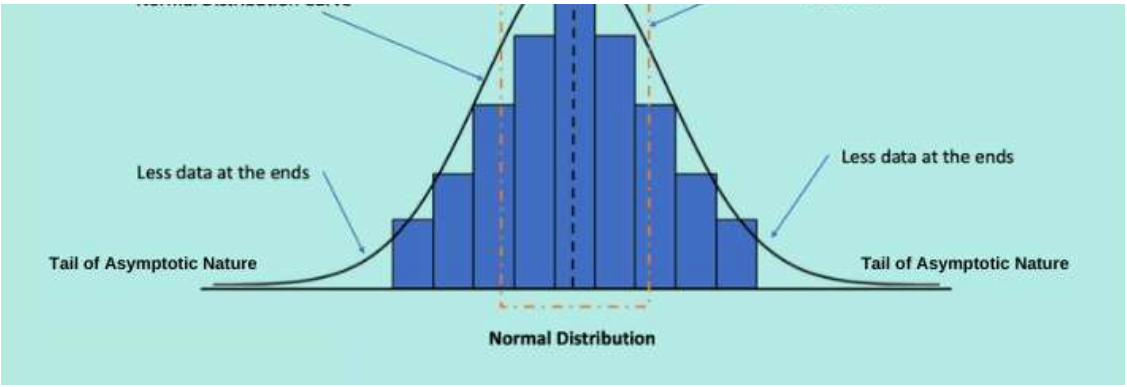
In [4]:
```
# Generate random numbers from the Gaussian distribution
num_samples = 1000
gaussian_data = np.random.normal(loc = 0.5, scale = 0.2, size = num_samples)
```

In [19]:
```
# Plot the histogram of the generated data
plt.figure(figsize=(8, 6))
sns.histplot(gaussian_data, bins=30, kde = True, color='orange', alpha=0.7)
plt.title('Histogram of Gaussian Distribution')
plt.xlabel('Value')
plt.ylabel('Density')
plt.show()
```



Common Properties of the Normal Distribution

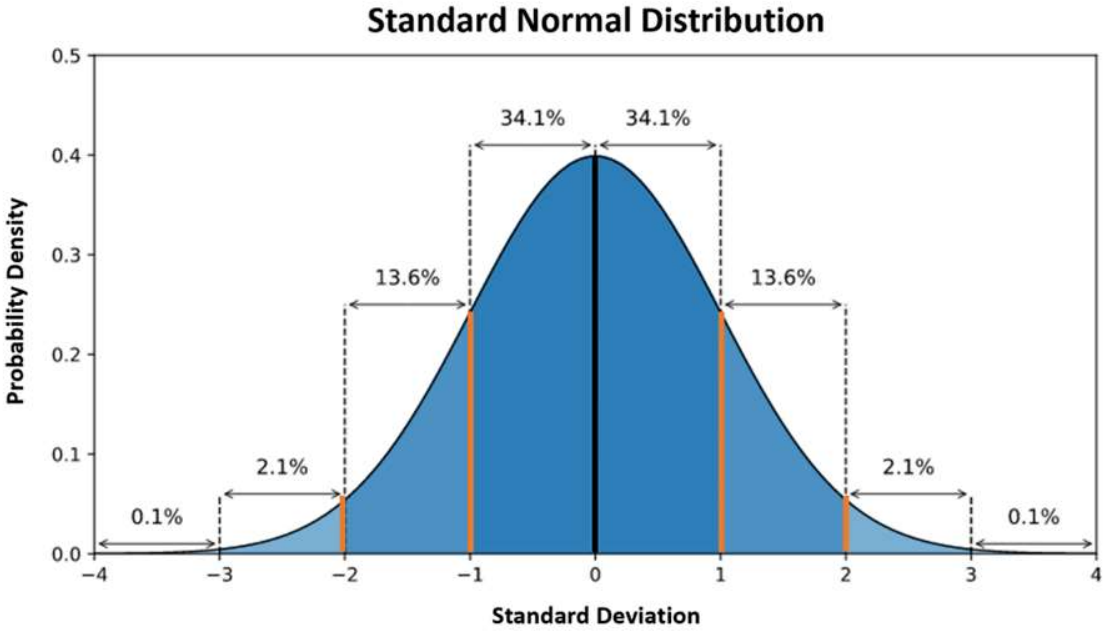
- **Unimodal and Symmetric**: A normal distribution has a single peak (unimodal) and is symmetric around its mean.
- **Equal Measures of Central Tendency**: The mean, median, and mode are all equal in a normal distribution.
- **Spread and Standard Deviation**: The standard deviation measures the spread of data. Roughly half of the population falls below the mean, and the other half lies above it.
- **Asymptotic Tails**: The tails of a normal distribution extend infinitely in both directions, approaching the x-axis.
- **Area Under the Graph**: In a normal distribution, the total area under the graph represents the total probability, and it always sums up to 1 (for every probability density function (PDF))
- **68-95-99.7 Rule**: The empirical rule states that about 68% of the data falls within one standard deviation of the mean, about 95% falls within two standard deviations, and about 99.7% falls within three standard deviations.



Standard Normal Variate / Standard Normal Distribution (Z)

It is a Standardized form of Normal Distribution, where the mean is zero and the standard deviation is 1. This distribution is also known as the Z-distribution.

Z ~ N(0, 1)



Standard Scores

A standard score represents the number of standard deviations above or below the mean that a specific observation falls.

For example, a standard score of -2.65 indicates that the observation is 2.65 standard deviations below the mean. On the other hand, a positive score represents a value above the average. The mean has a Z-score of 0.

Why do we need Standard Scores?

The standard score/ z-score. is a very useful statistic for two main reasons:

- **Probability Calculation**: It allows us to calculate the probability of a score occurring within a normal distribution.
- **Comparison Across Distributions**: It enables us to compare two scores that come from different normal distributions

Note: Standardizing your data before applying machine learning or deep learning models is a recommended practice. It ensures that features are on a common scale, which helps algorithms perform better and converge faster during training to make robust model and improve its ability to generalize across different datasets.

How to Calculate Z-scores?

Mathematically, to calculate the Z-score of a data point in a normal distribution:

$$Z = \frac{X - \mu}{\sigma}$$

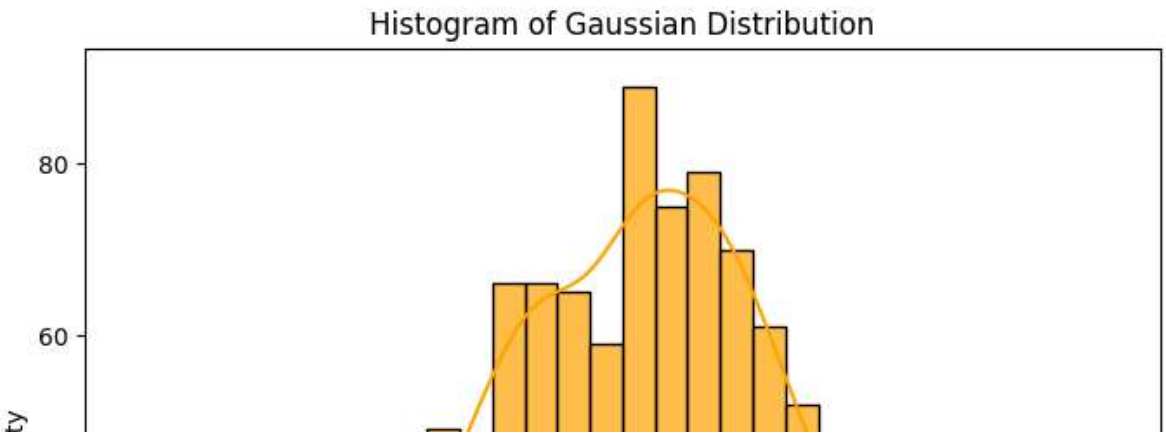
Where:

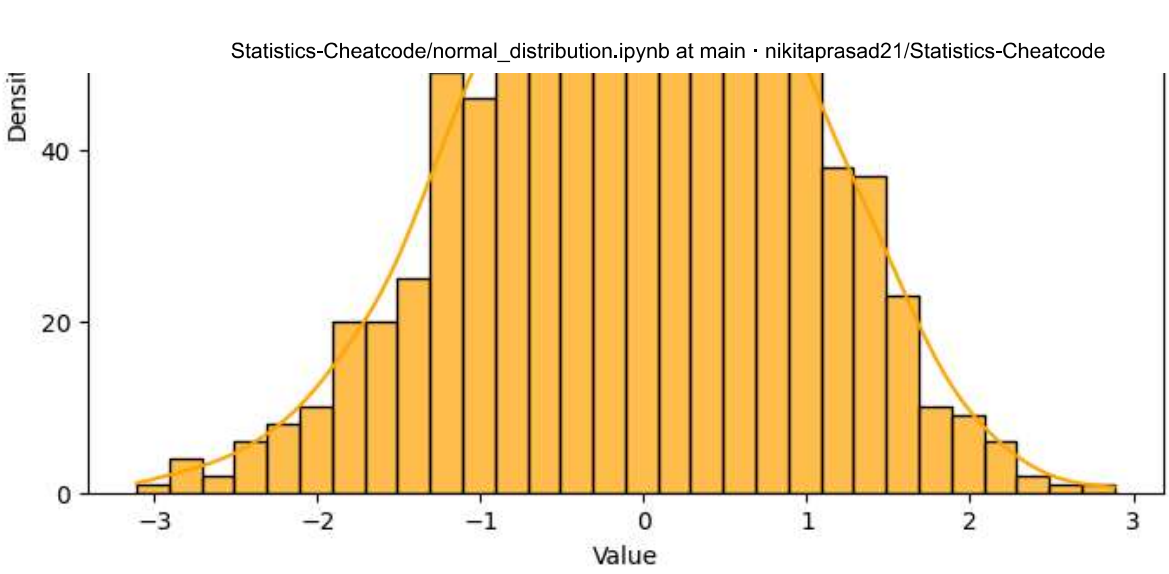
- Z is the Z-score.
- X is the value of the data point.
- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.

In [14]:
```
# Set the mean and standard deviation of the Gaussian distribution
mean = 0
std_dev = 1
```

In [15]:
```
# Generate random numbers from the Gaussian distribution
num_samples = 1000
standardized_data = np.random.normal(mean, std_dev, num_samples)
```

In [18]:
```
# Plot the histogram of the generated data
plt.figure(figsize=(8, 6))
sns.histplot(standardized_data, bins=30, kde=True, color='orange', alpha=0.7)
plt.title('Histogram of Gaussian Distribution')
plt.xlabel('Value')
plt.ylabel('Density')
plt.show()
```

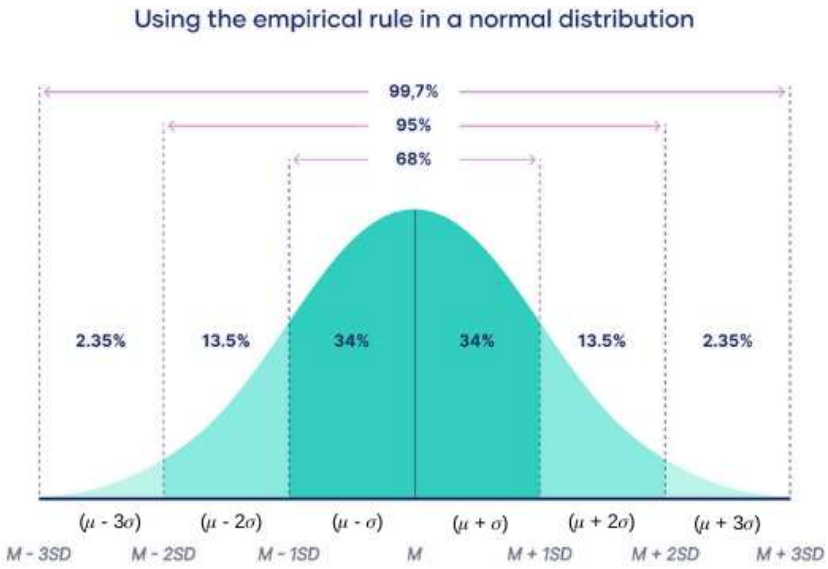




Emperical Rule

It states that approximately 68% data falls within It states that for a normal distribution:

- Approximately **68%** of the data falls within **one standard deviation (σ) of the mean (μ).**
- Approximately **95%** of the data falls within **two standard deviations (2σ) of the mean (μ).**
- Approximately **99.7%** of the data falls within **three standard deviations (3σ) of the mean (μ).**



Use of Normal Distribution in Data Science and ML

- **Statistical Modeling**: Normal distribution is often assumed in statistical modeling due to its simplicity and wide applicability, making it a fundamental concept in regression analysis, hypothesis testing, and ANOVA.
- **Feature Engineering**: Identifying and transforming features to follow a normal distribution can improve the performance of linear models and algorithms like linear regression and logistic regression.
- **Outlier Detection**: Normal distribution-based methods, such as Z-score or Mahalanobis distance, are commonly used for detecting outliers in datasets.

- **Central Limit Theorem**: It forms the basis of the Central Limit Theorem, which states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, allowing for the estimation of population parameters.
- **Error Distribution**: In many machine learning algorithms, such as linear regression and Gaussian Naive Bayes, the assumption of normally distributed errors is made for accurate parameter estimation and inference.
- **Residual Analysis**: Checking the normality of residuals is essential in assessing the goodness-of-fit of regression models and ensuring the validity of statistical inferences.
- **Anomaly Detection**: Gaussian Mixture Models (GMMs) are employed for anomaly detection by modeling normal behavior and identifying deviations from it.
- **Classification**: In Bayesian classifiers like Naive Bayes, normal distribution is assumed for continuous features, allowing for efficient and accurate classification.
- **Model Evaluation**: Normality tests, such as Shapiro-Wilk test or Kolmogorov-Smirnov test, are used to assess the normality of data distributions, which is crucial for selecting appropriate statistical tests and evaluating model assumptions.

How to find if a given distribution is normal or not?

Here are a few common methods:

- **Visual Inspection**:
  - **Histogram**: A bell-shaped curve with symmetrical data around the mean indicates normality.
  - **Q-Q Plot**: A straight line suggests normal distribution.
- **Statistical Tests**:
  - **Shapiro-Wilk, Anderson-Darling, and Kolmogorov-Smirnov** tests compare data to expected normal distribution.
  - Low p-values (< 0.05) indicate non-normality.
- **Summary Statistics**:
  - **Skewness**: Normal distribution has skewness of 0.
  - **Kurtosis**: Normal distribution has kurtosis of 3.

> Stay tuned for more Statistical Concepts and Don't forget to **Star** this Github Repository for more such contents and consider sharing with others.

In [ ]: